Large Language Models Support Radiologists in Detecting Overlooked Crohn's Disease in Urolithiasis CT reports

Los modelos de lenguaje de gran tamaño ayudan a los radiólogos a detectar casos inadvertidos de enfermedad de Crohn en informes de TC realizados por urolitiasis

Teodoro Martín-Noguerol, MD[1]; Pilar López-Úbeda PhD[2]; Manuel García-Ferrer, MD[3]; Alba Ruiz-Vinuesa, BS[4]; Antonio Luna, MD, PhD[1]

[1] MRI unit, Radiology department. HT Médica, Jaén, Spain
[2] NLP department. HT Médica. Jaén, Spain
[3] Radiology Department. HT Médica, Sevilla, Spain
[4.] Escuela de Ingeniería de Fuenlabrada, Universidad Rey Juan Carlos, Madrid, Spain
t.martin.f@htime.org, p.lopez@htmedica.com, m.garcia@htmedica.com, a.ruizv.2018@alumnos.urjc.es, aluna70@htmedica.com

Abstract: Crohn's disease and urolithiasis coexist in a significant number of patients. This study evaluates the ability of Large Language Models (LLMs) and rule-based Natural Language Processing (NLP) tools to automatically detect terms related to both conditions in abdominal Computed Tomography (CT) reports. A random sample of 290 reports in Spanish was selected from a total of 163,189 abdominopelvic CT studies performed between 2019 and 2025. They were categorized into four groups: Crohn's disease, urolithiasis, both conditions, or neither. A rule-based method using the Unified Medical Language System (UMLS) was developed as a baseline. Additionally, four LLMs (Llama-3-70B, Gemini 2.5 Flash, MedGemma, and Cohere Command A) were evaluated using prompting techniques. To assess generalizability, a separate prospective validation was conducted with 50 independent reports. Cohere Command A showed the best overall performance, achieving an accuracy of 0.852 and a macro-F1 score of 0.849. While the UMLS-based rule method performed reasonably well, it was consistently outperformed by the advanced language models. Prospective validation revealed that the Cohere model induced diagnostic revisions in 41% of ambiguous cases, improving detection of combined Crohn's and urolithiasis pathology. LLMs, particularly Cohere Command A, can extract relevant categories from CT reports and support diagnostic decision-making. Their integration into clinical workflows could enhance disease detection and consistency in radiology reporting, as demonstrated by their impact during prospective validation.
Keywords: Crohn's disease, urolithiasis, NLP, LLM

Resumen: La enfermedad de Crohn y la urolitiasis coexisten en un número significativo de pacientes. Este estudio evalúa la capacidad de los modelos de lenguaje a gran escala (LLM) y de las herramientas de procesamiento del lenguaje natural (PLN) basadas en reglas para detectar automáticamente términos relacionados con ambas afecciones en los informes de tomografía computarizada (TC) abdominal. Se seleccionó una muestra aleatoria de 290 informes en español de un total de 163.189 estudios de TC abdominopélvica realizados entre 2019 y 2025. Se clasificaron en cuatro grupos: enfermedad de Crohn, urolitiasis, ambas afecciones o ninguna de ellas. Se desarrolló un método basado en reglas que utiliza el Sistema Unificado de Lenguaje Médico (UMLS) como referencia. Además, se evaluaron cuatro LLMs (Llama-3-70B, Gemini 2.5 Flash, MedGemma y Cohere Command A) utilizando técnicas de prompting. Para evaluar la generalizabilidad, se llevó a cabo una validación prospectiva independiente con 50 informes. Cohere Command A mostró el mejor rendimiento general, alcanzando una

precisión de 0,852 y una puntuación macro-F1 de 0,849. Si bien el método basado en reglas del UMLS funcionó razonablemente bien, los modelos de lenguaje avanzados lo superaron. La validación prospectiva reveló que el modelo Cohere indujo revisiones diagnósticas en el 41% de los casos ambiguos, mejorando la detección de la patología combinada de la enfermedad de Crohn y la urolitiasis. Los LLMs, en particular Cohere Command A, pueden extraer categorías relevantes de los informes de TC y apoyar la toma de decisiones diagnósticas. Su integración en los flujos de trabajo clínicos podría mejorar la detección de enfermedades y la coherencia en los informes radiológicos.
Palabras clave: enfermedad de Crohn, urolitiasis, PLN, LLM

1 *Introduction*

The association between Crohn's disease and urolithiasis is well-known in the scientific literature (Miyajima et al., 2021; van Sommeren et al., 2014). Diverse studies have established an incidence of nephrolithiasis in patients with Crohn's disease of around 12% (Gkentzis et al., 2016; Olpin et al., 2017). Inflammation of the terminal ileum results in fatty malabsorption, which subsequently triggers increased oxalate reabsorption due to calcium binding with fatty acids. Elevated blood oxalate levels lead to a higher prevalence of urolithiasis in Crohn's disease patients compared to the general population (Torricelli et al., 2021). However, this specific link between the two conditions can sometimes be overlooked in routine clinical and radiological evaluations. While Computed Tomography (CT) radiological reports might detail bowel signs associated with Crohn's disease detection, correlation of urolithiasis in patients with bowel findings might be absent or undervalued. Furthermore, it is not uncommon that radiologists consider the evaluation of inflammatory bowel signs when examining CT scans targeted to evaluate the kidneys, ureters, and bladder (KUB) for ruling out urolithiasis (Sarofim et al., 2016).

In such contexts, apart from the expertise and familiarity of radiologists with this association, text-based Artificial Intelligence (AI) tools can assist both novice and seasoned radiologists (Hosny et al., 2018; Noguerol et al., 2019). Natural Language Processing (NLP) solutions, including the recent Large Language Models (LLMs), can extract relevant information from the radiology report automatically to warn radiologists in real time about the potential association of the findings described in their radiology report (López-Úbeda et al., 2024, 2022). With this backdrop, we aim to test and compare rule-based NLP algorithms and LLMs' performance as a clinical decision support tool for radiologists. This tool would guide and alert them in real-time about possible correlations between Crohn's disease and urolithiasis as they draft their radiology reports. The primary objective of this NLP solution is to improve the identification of links between Crohn's disease manifestations and urolithiasis indicators in radiology reports, ultimately enhancing patient care and overall clinical management.

2 *Material and Methods*

2.1 Study Design

Due to the retrospective nature of the study, which was based on anonymized radiology reports, the local Ethics Committee waived the need for ethical approval and informed consent. This study analyzed unstructured abdominopelvic CT reports, dated from January 2019 to January 2025. A total of 163,189 CT reports were retrieved from our Radiology Information System (RIS) as the initial dataset. All reports were anonymized and typically organized into sections such as clinical information, radiological findings, and conclusions.

From this dataset, a stratified random sample of 290 reports was selected to ensure adequate representation across four predefined clinical categories, based on the presence of specific pathologies mentioned in the radiological text:

- Crohn group: presence of Crohn's disease,
- Urolithiasis group: presence of urolithiasis,
- Crohn & urolithiasis group: presence of both Crohn's disease and urolithiasis,
- Control group: absence of both pathologies.

The 290 radiology reports were written in Spanish and had a mean length of 235 ± 97 words (median: 226 words; interquartile range: 166–287 words; range: 52–766 words), remaining well below the context window limits of all evaluated LLMs.

All reports in the sample were manually annotated by an internal board-certified radiologist with over 15 years of experience.[1] This sampling approach preserved the diagnostic diversity of the overall dataset while enabling a practical, reproducible, and clinically relevant analysis. Figure 1 shows the distribution of groups annotated by the radiologist in the 290 CT reports. The urolithiasis group includes 102 cases, the Control group has 72 cases, the Crohn & urolithiasis group contains 61 cases, and the Crohn group consists of 55 cases. These numbers reflect the absolute distribution of the study population, which is also represented in the percentage distribution shown in the figure.

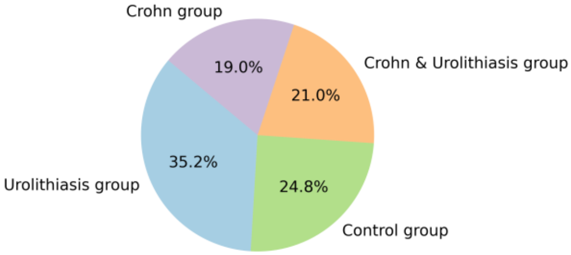

Figure 1. Percentage distribution of 290 annotated CT reports classified into four clinical categories: Crohn’s disease, urolithiasis, both conditions (Crohn & urolithiasis), and no pathology (Control).

The study sample consisted of patients with a wide age range, with ages spanning from 8 to 92 years (mean ± SD: 56 ± 16.6 years). The distribution of sex was balanced, with 51% male and 49% female participants. Furthermore, these reports have been written by 66 different radiologists at 17 different centers within our institution.

2.2 Experimental Setup

To evaluate the automatic classification of CT reports into the four predefined clinical categories, we compared five different methods, covering both rule-based and LLM-based approaches.

2.2.1 Rule-Based Method

This baseline method classifies radiology reports based on the presence of predefined clinical concepts from the Unified Medical Language System (UMLS). The approach uses exact string matching to identify medical terms explicitly associated with Crohn’s disease and urolithiasis within the report text. For the Crohn group, we used the UMLS concept for Crohn’s disease (CUI: C0010346) as the primary entry point. To improve sensitivity, we expanded the search to include its 36 narrower concepts, encompassing related terms such as regional enteritis and terminal ileitis. Similarly, for the Urolithiasis group, we selected the concept for urolithiasis (CUI: C0451641) and included 38 narrower concepts, such as renal calculi, nephrolithiasis, and ureterolithiasis. This expansion ensures the method captures clinically relevant subtypes and terminology variations that may appear across reports.

To improve the accuracy of concept detection, the NegEX-MES negation detection system was employed to determine whether the identified clinical terms appeared in a negated context within the text. Only terms not flagged as negated were considered valid for classification. Additionally, to avoid misclassification caused by terms related to other anatomical regions, reports mentioning concepts associated with the gallbladder, biliary tract, common bile duct (choledochus), hepatic, cystic, or pancreatic regions were excluded when identifying urolithiasis. This filtering step was crucial to prevent false positives from terms like "lithiasis" that could refer to calculi outside the urinary system. A report was assigned to a clinical group if it contained at least one matched concept. Reports containing concepts for both Crohn’s disease and urolithiasis were classified into the Crohn's & lithiasis group. If no target concepts were found, the report was labeled as control.

2.2.2 Llama

Llama 3 is a state-of-the-art open-source LLM developed by Meta AI. In this study, we used the Llama-3-70B model to classify each radiology report. The model received only the relevant textual sections: clinical information, radiological findings, and conclusions, followed by an instruction to assign one of the four clinical categories as shown in Appendix A (the prompt presented in the manuscript has been translated into English solely to improve readability and facilitate understanding). Inference was performed in a zero-shot manner, with no fine-tuning.

2.2.3 Gemini

Gemini is Google DeepMind’s LLM platform. We used the Gemini 2.5 Flash version to classify

[1] The corpus will be available upon request to the author.

reports using the same standardized prompt that was applied for Llama, enabling consistent evaluation across models.

2.2.4 MedGemma

MedGemma is a collection of Gemma 3 variants trained for medical text and image comprehension. We used MedGemma to classify reports using the same standardized prompt that was applied for the other LLMs, enabling a consistent comparison across models.

2.2.5 Cohere Command A

Command A is an LLM developed by Cohere, available through Oracle Cloud Infrastructure. In this study, we used the Command A a.03-2025 version. The model was prompted with the same content used for the other LLMs.

2.3 Implementation and validation in the radiology department workflow

The NLP-based algorithm was developed using the Python programming language for the automated detection of keywords related to Crohn's disease and urolithiasis within radiology reports. This system was integrated into the daily workflow of our radiology department, where it processes each patient's current abdominopelvic CT report as input. The NLP algorithm identifies relevant keywords associated with the four predefined clinical categories (Crohn group, urolithiasis group, Crohn & urolithiasis group, and Control group). It classifies the report accordingly, while the LLMs provide the group directly.

To evaluate the integrated system's performance in a real-world setting, we selected the best of the tested systems (rule-based NLP, Llama, Gemini, MedGemma or Cohere). A validation study was conducted using a random sample of 50 reports from the initial dataset, all of which were annotated as belonging to the "Crohn & urolithiasis group" but were blinded to the radiologist. These reports were not used in any of the previous NLP algorithm evaluations, ensuring that the validation relied on completely unseen data. The 50 anonymized studies were independently reviewed by an external board-certified radiologist with 8 years of experience. This radiologist analyzed the imaging findings and generated their corresponding radiology reports for each case. Additionally, Appendix B contains implementation details.

2.4 Statistical Analysis

To assess the performance of each classification method, we compared the predicted labels against the gold-standard annotations. We selected macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and accuracy, which provide a global assessment of model performance across all classes. Moreover, precision, recall, and F1-score were calculated separately for each group (Crohn, urolithiasis, Crohn & urolithiasis, and Control) to examine the performance at a per-class level. To further evaluate the model's discrimination capabilities, receiver operating characteristic (ROC) curves were generated for each method, and the corresponding area under the curve (AUC) values were reported. Finally, a confusion matrix was presented as the best-performing method to illustrate the distribution of true and predicted labels across all groups.

3 Results

Table 1 shows the evaluation results of the methods evaluated in this study. Among the five methods evaluated, Cohere command A stood out with the best overall performance, achieving the highest scores in all metrics, including accuracy (0.852) and macro-F1 (0.849). Gemini also showed solid and balanced performance, only slightly below Cohere. The rule-based method based on UMLS concepts achieved competitive results, ranking above both Llama and MedGemma despite its relatively simple design. Llama showed moderate performance, with an accuracy of 0.628 and a macro-F1 of 0.608. MedGemma obtained the lowest overall performance, with an accuracy of 0.514 and a macro-F1 of 0.479.

Precision, recall, and F1-score for each patient group: a) Crohn group, b) Urolithiasis group, c) Crohn & urolithiasis group, and d) Control group, as well as the results for each method compared are shown in Figure 2. Overall, and consistent with the global results presented in the previous table, the Cohere method achieves the best performance across all categories, particularly in the Crohn group. In contrast, the Llama model shows low recall and F1-score when classifying the Control group.

Method	Macro-precision	Macro-recall	Macro-F1	Accuracy
Rule-based	0.793	0.786	0.775	0.793
Llama	0.669	0.644	0.608	0.628
Gemini	0.809	0.808	0.807	0.814
MedGemmaa	0.570	0.509	0.479	0.514
Cohere	**0.847**	**0.853**	**0.849**	**0.852**

Table 1. Performance comparison of five classification methods across defined groups. The best results for each metric are in bold.

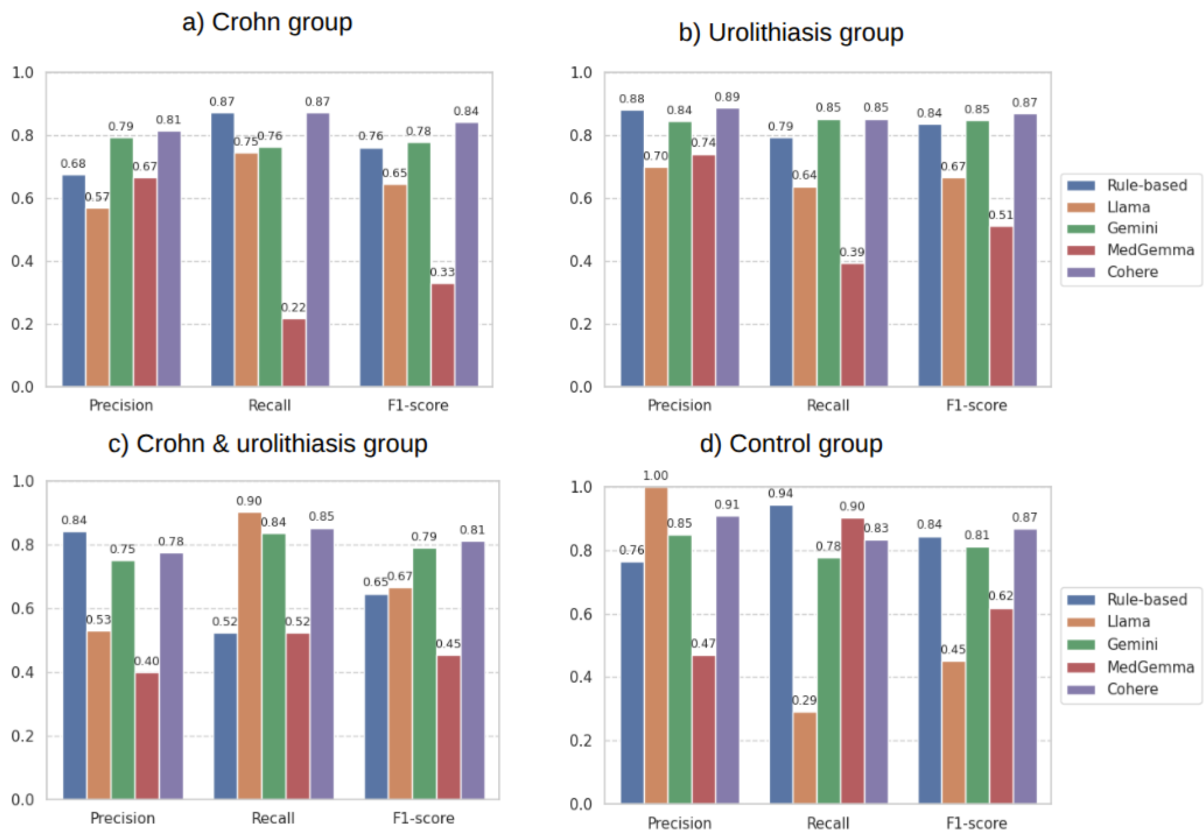

Figure 2. Precision, recall, and F1-score for each patient group and classification method. The metrics are shown for the Crohn group, Urolithiasis group, Crohn & Urolithiasis group, and Control group, highlighting the comparative performance of the evaluated methods.

Additionally, Figure 3 presents the ROC (Receiver Operating Characteristic) for five different classification models. The Area Under the Curve (AUC) metric is used to evaluate their performance, with higher AUC values indicating better discrimination between positive and negative classes. Among the models, Cohere achieves the highest AUC (0.90), demonstrating superior classification ability, followed by Gemini (0.87) and the rule-based model (0.86). Llama and MedGemma show the lowest performance with an AUC of 0.76 and 0.67 respectively.

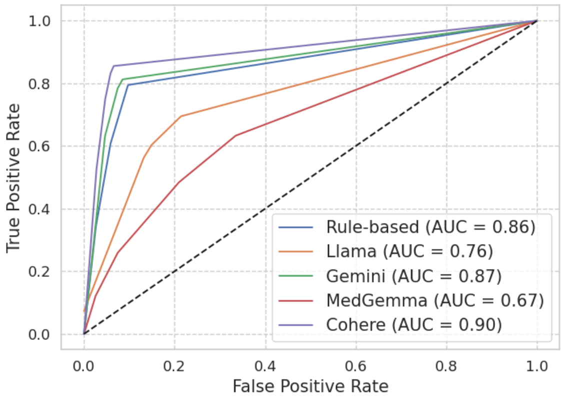

Figure 3. Receiver operating characteristic curve using rule-based, Llama, Gemini, MedGemma and Cohere.

Qualitative inspection of misclassified reports revealed that some challenging cases contained explicit evidence of urolithiasis but only indirect indicators of Crohn's disease, such as chronic inflammatory changes, bowel stenosis, postoperative findings, and perianal fistulous tracts. In these reports, simpler approaches often identified only one of the two conditions, whereas the best-performing LLMs were more frequently able to infer their coexistence from the broader clinical context.

Finally, considering the method that yielded the best results (Cohere), its confusion matrix is shown in Figure 4. The Cohere model has generally good performance, especially in classifying the urolithiasis and Crohn groups. The errors are relatively low, although there is some confusion between the Crohn & urolithiasis and Control groups, as well as between the Crohn and Urolithiasis groups.

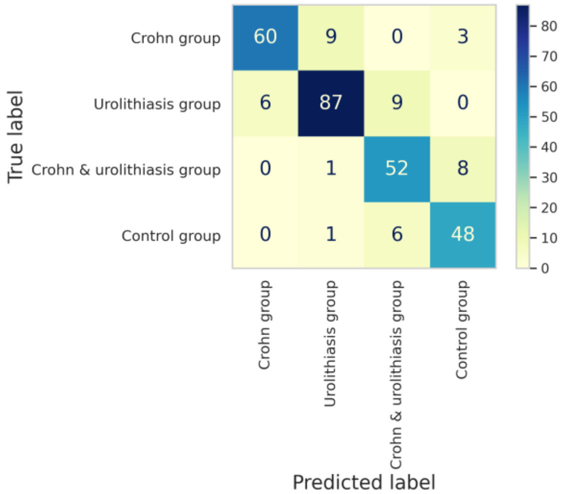

Figure 4. Confusion matrix of the best-performingr method (Cohere), illustrating the classification results across the four groups.

Validation in the radiology department

Based on prior performance evaluations, we deployed Cohere as the most accurate system for direct classification of radiology reports into four predefined clinical categories. To validate the system, a radiologist reviewed an additional set of 50 abdominal CT reports. Initially, the radiologist correctly classified 28 out of 50 cases (56%) as involving both Crohn's disease and urolithiasis. In the remaining 22 cases (44%), the radiologist identified only urolithiasis-related findings.

For these 22 reports, the Cohere-based system issued classification alerts, suggesting that the CT studies may also involve Crohn-related features. The radiologist re-evaluated the reports considering the system's classifications and subsequently revised the diagnosis in 9 out of the 22 cases, updating the final report to reflect the dual pathology. In the remaining 13 cases, the original classification was maintained. These corresponded to patients with a previous history of Crohn's disease, but no active disease imaging findings at the current CT studies.

4 *Discussion*

We have developed a rule-based NLP algorithm and tested and compared it with other commercially available LLMs for extracting terms related to Crohn's disease at abdominal-pelvic CT reports and urolithiasis to assist radiologists in their workflow to remind them of the association between both topics. The highest accuracy for this task was obtained with the Cohere LLM (0.852), while the lowest accuracy was obtained with

Llama. The Cohere LLM has been integrated into our RIS as part of our radiology reporting workflow assisting radiologists during the report generation process in terms of advising them about considering the possibility of a relationship between both entities or warning about seeking findings related to Crohn's disease in patients with urolithiasis and vice versa, finding a positive impact on the way these studies are reported. This validation phase demonstrated that the system influenced the radiology reporting process in (9/22) of cases overall, in which no urolithiasis was initially detected on the first radiology report draft. These results support the utility of our approach in enhancing diagnostic accuracy and clinical decision-making in radiology workflows.

To the best of our knowledge, this is the first approach that addresses the use of NLP tools, based on radiology reports, for helping radiologists with the association between Crohn's disease and urolithiasis. Despite the current trend in the use of NLP solutions for exploring and exploiting radiology reports, scarce studies have evaluated the potential role of NLP, in helping radiologists in the assessment of specific features of Crohn's disease or urolithiasis even separately, being usually focused on image-based AI algorithms (Javaid et al., 2022; Liu et al., 2022; Luo and Chong, 2020; Sleiman et al., 2022). Concerning the detection of urolithiasis using NLP, Jungmann et al automatically structured radiology reports based on RadLex terms and were able to detect positive terms related to urolithiasis, especially in patients with a previous history of stones and obstructive uropathy (Jungmann et al., 2020; Tsuji et al., 2021).

Concerning NLP's application to Crohn's disease, some researchers have employed Electronic Health Records (EHRs) to feed text-based algorithms (Montoto et al., 2022). Stidham et al used information from EHRs processed with NLP to identify the presence, activity, and status of extraintestinal manifestations of inflammatory bowel disease (Stidham et al., 2023). They found an association with arthritis, ocular, and psoriasiform manifestation; however, they did not mention potential relationships with urolithiasis. Ananthakrishnan et al. used EHRs for improving case definition of inflammatory bowel diseases, including Crohn's disease and ulcerative colitis (Ananthakrishnan et al., 2013). They found that including narrative concepts about both diseases previously identified using NLP, the accuracy of EHRs case for inflammatory bowel disease improved compared with models that use codified data alone. In the same line, Tong et al applied random forest and convolutional neural networks for differentiating between ulcerative colitis, Crohn's disease and intestinal tuberculosis based on the description of the endoscopic image from free text, obtaining high sensitivity and specificity especially for distinguishing between Crohn's disease and intestinal tuberculosis (Tong et al., 2020).

Our study has some limitations. Concerning the number and type of studies selected for the automatic evaluation of the system, a small cohort of 290 abdominopelvic CT reports was used. Additionally, all reports were annotated by a single experienced radiologist, and no inter-observer agreement analysis was performed. Multi-reader annotation was not feasible due to the substantial time and clinical resources required for expert labeling of radiology reports. We used UMLS to determine specific keywords. However, further studies will need to determine if the use of personalized keywords instead, based on local regular expressions, would determine results more aligned with our reality (Galperin et al., 2022). It should be noted that the rule-based tool used in this study is limited by a static vocabulary, and no fine-tuning was applied to the LLM, which was used in a zero-shot manner. Furthermore, the best-performing model (Cohere) is a proprietary system that provides limited explainability regarding its internal reasoning process. Although the tool is intended to support rather than replace radiologist decision-making, future work should investigate methods to improve transparency and explainability in clinical deployment. Regarding the terms excluded from the analysis, such as all related to cholelithiasis, we are aware that there is also a potential association described between lithiasis of biliary origin and inflammatory bowel disease (Chen et al., 2018; Fagagnini et al., 2017). Probably further studies applying our approach may also assist radiologists for these other couple of related entities. Furthermore, although the prospective validation included an independent radiologist who re-evaluated the CT studies, no formal adjudication process involving additional experts was performed to confirm the correctness of all AI-induced report revisions. For a more comprehensive assessment of the performance of LLMs for this task, a larger sample of studies should be evaluated by radiologists. Future enhancements to our algorithm could involve retrospective examinations of previous reports for a patient, leveraging LLMs to better inform current radiology requests from clinicians, and how this novel approach may change management or improve the clinical outcome of these patients. Finally, future work should explore the optimization of rule-based approaches, including the incorporation of local terminology and more sophisticated contextual rules, given the competitive performance achieved by the baseline rule-based model in this study.

5 Conclusion

The application of LLMs to accurately identify and associate Crohn's disease and urolithiasis within a single patient's radiology report is highly innovative. This approach can assist radiologists during the reporting process by helping ensure that early or advanced terminal ileum involvement is not overlooked in CT studies of patients with urolithiasis. Furthermore, it supports the identification of the likely origin and nature of urolithiasis in patients diagnosed with Crohn's disease.

Bibliography

Ananthakrishnan, A.N., Cai, T., Savova, G., Cheng, S.-C., Chen, P., Perez, R.G., Gainer, V.S., Murphy, S.N., Szolovits, P., Xia, Z., others, 2013. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm. Bowel Dis. 19, 1411–1420.

Chen, C.-H., Lin, C.-L., Kao, C.-H., 2018. Association between inflammatory bowel disease and cholelithiasis: a nationwide population-based cohort study. Int. J. Environ. Res. Public Health 15, 513.

Fagagnini, S., Heinrich, H., Rossel, J.-B., Biedermann, L., Frei, P., Zeitz, J., Spalinger, M., Battegay, E., Zimmerli, L., Vavricka, S.R., others, 2017. Risk factors for gallstones and kidney stones in a cohort of patients with inflammatory bowel diseases. PLoS One 12, e0185193.

Galperin, R., Schnapp, S., Elhadad, M., 2022. Cross-lingual UMLS named entity linking using UMLS dictionary fine-tuning, in: Findings of the Association for Computational Linguistics: ACL 2022. pp. 3380–3390.

Gkentzis, A., Kimuli, M., Cartledge, J., Traxer, O., Biyani, C.S., 2016. Urolithiasis in inflammatory bowel disease and bariatric surgery. World J. Nephrol. 5, 538.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.W.L., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18, 500–510.

Javaid, A., Shahab, O., Adorno, W., Fernandes, P., May, E., Syed, S., 2022. Machine learning predictive outcomes modeling in inflammatory bowel diseases. Inflamm. Bowel Dis. 28, 819–829.

Jungmann, F., Kämpgen, B., Mildenberger, Philipp, Tsaur, I., Jorg, T., Düber, C., Mildenberger, Peter, Kloeckner, R., 2020. Towards data-driven medical imaging using natural language processing in patients with suspected urolithiasis. Int. J. Med. Inform. 137, 104106.

Liu, Y.-Y., Huang, Z.-H., Huang, K.-W., 2022. Deep learning model for computer-aided diagnosis of urolithiasis detection from kidney–ureter–bladder images. Bioengineering 9, 811.

López-Úbeda, P., Martín-Noguerol, T., Escartín, J., Luna, A., 2024. Role of natural language processing in automatic detection of unexpected findings in radiology reports: a comparative study of RoBERTa, CNN, and ChatGPT. Acad. Radiol. 31, 4833–4842.

López-Úbeda, P., Martín-Noguerol, T., Juluru, K., Luna, A., 2022. Natural language processing in radiology: update on clinical applications. Journal of the American College of Radiology 19, 1271–1285.

Luo, J.W., Chong, J.J.R., 2020. Review of natural language processing in radiology. Neuroimaging Clinics 30, 447–458.

Miyajima, S., Ishii, T., Watanabe, M., Ueki, T., Tanaka, M., 2021. Risk factors for urolithiasis in patients with Crohn's disease. International Journal of Urology 28, 220–224.

Montoto, C., Gisbert, J.P., Guerra, I., Plaza, R., Pajares Villarroya, R., Moreno Almazán, L., López Martín, M.D.C., Dom\'\inguez Antonaya, M., Vera Mendoza, I., Aparicio, J., others, 2022. Evaluation of natural language processing for the identification of Crohn disease–related variables in Spanish electronic health records: a validation study for the PREMONITION-CD Project. JMIR Med. Inform. 10, e30345.

Noguerol, T.M., Paulano-Godino, F., Mart\'\in-Valdivia, M.T., Menias, C.O., Luna, A., 2019. Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. Journal of the American College of Radiology 16, 1239–1247.

Olpin, J.D., Sjoberg, B.P., Stilwill, S.E., Jensen, L.E., Rezvani, M., Shaaban, A.M., 2017. Beyond the bowel: extraintestinal manifestations of inflammatory bowel disease. Radiographics 37, 1135–1160.

Sarofim, M., Teo, A., Wilson, R., 2016. Management of alternative pathology detected using CT KUB in suspected ureteric colic. International Journal of Surgery 32, 179–182.

Sleiman, J., Chirra, P., Gandhi, N.S., Baker, M.E., Lu, C., Gordon, I.O., Viswanath, S.E., Rieder, F., Therapy, S., Consortium, A.-F.R. (STAR), 2022. Crohn's disease related strictures in cross-sectional imaging: More than meets the eye? United European Gastroenterol. J. 10, 1167–1178.

Stidham, R.W., Yu, D., Zhao, X., Bishu, S., Rice, M., Bourque, C., Vydiswaran, V.V.G., 2023. Identifying the presence, activity, and status of extraintestinal manifestations of inflammatory bowel disease using natural language processing of clinical notes. Inflamm. Bowel Dis. 29, 503–510.

Tong, Y., Lu, K., Yang, Y., Li, J., Lin, Y., Wu, D., Yang, A., Li, Y., Yu, S., Qian, J., 2020. Can natural language processing help differentiate inflammatory intestinal diseases in China? Models applying random forest and convolutional neural network approaches. BMC Med. Inform. Decis. Mak. 20, 248.

Torricelli, F.C., Reichard, C., Monga, M., 2021. Urolithiasis in complicated inflammatory bowel disease: a comprehensive analysis of urine profile and stone composition. Int. Urol. Nephrol. 53, 205–209.

Tsuji, S., Wen, A., Takahashi, N., Zhang, H., Ogasawara, K., Jiang, G., 2021. Developing a RadLex-based named entity recognition tool for mining textual radiology reports: development and performance evaluation study. J. Med. Internet Res. 23, e25378.

van Sommeren, S., Janse, M., Karjalainen, J., Fehrmann, R., Franke, L., Fu, J., Weersma, R.K., 2014. Extraintestinal manifestations and complications in inflammatory bowel disease: from shared genetics to shared biological pathways. Inflamm. Bowel Dis. 20, 987–994.

A Example prompt used to interact with LLMs

It is well documented that patients with Crohn's disease have an increased risk of developing urolithiasis.
You will now be provided with a radiology report.
Your task is to classify the report into one of the following four clinical groups, based on the presence of terms related to specific conditions, as defined in the UMLS:
- Crohn group: presence of terms related to Crohn's disease.
- Urolithiasis group: presence of terms related to urolithiasis.
- Crohn & urolithiasis group: presence of terms related to both Crohn's disease and urolithiasis.
- Control group: absence of terms related to either condition.

Please respond only with the name of the group.
Analyze the following report:
CLINICAL INFORMATION: Abdominal pain. AP Crohn's disease.
FINDINGS: Postoperative changes in the ileocecal region are observed. Significant wall thickening of the distal neoileum was observed with slight rarefaction of the surrounding fat, the presence of some adjacent fluid bands, and a small amount of free fluid in the pelvis. Regarding the aforementioned clinical history of Crohn's disease, the findings described suggest inflammatory activity in the distal ileum with secondary subocclusive symptoms. Right kidney stones. No dilatation of the excretory system.
CONCLUSION: Findings suggestive of intestinal inflammatory activity in the terminal neoileum with secondary subocclusive symptoms. Right kidney stones.

B Implementation details

Inference parameters:

All models were evaluated in a zero-shot setting without additional fine-tuning or retrieval augmentation.

To maximize reproducibility, all large language models were evaluated using deterministic generation settings. Temperature was fixed at 0.0 whenever supported by the platform. Top-p was set to 1.0, and no frequency or presence of penalties were applied. The maximum output length was limited to 10 tokens, as the expected output consisted exclusively of one of the four predefined categories.

Software environment:

The NLP and LLM evaluation pipeline were implemented in Python v3.12. Report processing, prompting, inference, and result aggregation were performed through custom Python scripts on a machine equipped with an NVIDIA GeForce™ RTX 5060 GPU.

C Confusion matrices